건강 스트림 데이터의 다차원적 분석을 위한 저장 구조

신혜원, 임윤선, 김 명
이화여자대학교 컴퓨터학과
e-mail:{speggy, lys96, mkim}@ewha.ac.kr

A Storage Scheme of Health Data Stream for Multidimensional Analysis

Hea Won Shin, Yoonsun Lim, Myung Kim
Department of Computer Science and Engineering,
Ewha Womans University

요 약

유비쿼터스 의료 기술이 본격화되면서 센서 네트워크를 통해 환자의 건강 관련 데이터 스트림을 수집하여 위험상황을 탐지하고 지속적인 건강 상태를 모니터링할 수 있게 되었다. 그러나 방대한 양의 스트림 데이터로부터 의미 있는 데이터를 효과적으로 찾아내기 위해서는 실시간으로 데이터의 갱신과 집계 연산이 가능해야 하고 데이터의 압축이 효율적으로 처리될 수 있는 다차원 저장구조가 필요하다. 기존의 다차원 데이터 분석 도구인 OLAP 큐브 저장구조는 실시간 업데이트가 힘들고, 스트림 데이터 저장 구조인 DSMS들은 다차원 데이터 분석이 용이하지 않다. 이에 본 연구에서는 건강 스트림 데이터의 특징과 질의를 분석하고, 이러한 스트림 데이터에 적합한 저장구조의 요건을 제시하였다. 또한 점진적 갱신이 가능하고, 대용량 데이터를 시간 차원으로 압축, 삭제하기 용이하며 실시간에 분석 데이터 구축이 가능한 저장구조를 제안하고 그 효율성을 보였다.

1. 서론

유비쿼터스 컴퓨팅 환경이 구축되면서 센서를 통해 사용자의 건강 상태를 실시간으로 모니터링하는 서비스들이 본격화되었다. 이러한 건강 관리 시스템은 환자의 상태를 모니터링하면서 건강 정보 스트림 데이터를 발생시킨다. 이와 같은 스트림 데이터를 신속하게 분석하고 체계적으로 관리하면 환자의 위험 상황을 탐지하거나, 예외 상황의 패턴을 생성할 수도 있고, 환자의 진료에 필요한 정보를 추출해 낼 수도 있다.

스트림 데이터는 실시간에 지속적으로 발생되며 그 양이 방대하기 때문에 용도에 맞게 필터링하고 신속하게 저장, 관리해야 한다. 따라서 건강 관련 스트림 데이터의 특징을 파악하여 다차원 분석이 신속하게 이루어질 수 있도록 해야 하고, 사용자가 원하는 질의의 유형을 파악하여 그것을 반영할 수 있도록 수집된 데이터를 저장해야 한다. 또한 고차원 희박 데이터에 대한 처리, 다차원 범위 질의 처리, 실시간 업데이트가 가능해야 하며, 스트림 형태의 예외 데이터를 현실성 있는 공간에 저장하고, 과거의 데이터를 적절히 압축 저장함으로써 현재 데이터와의 비교 분석이 가능하도록 해야 한다.

최근 수년간 샘플링, 히스토그램, 웨이블릿 등을 이용한 여러 데이터 압축이나 시놉시스 구조, DSMS(Data Stream Management System)[1]와 같은 시스템의 개발을 통하여 스트림 데이터들을 현실적으로 적용하기 위한 연구가 진행되었다. 그러나 이들은 대부분 스트림 데이터를 1차원적으로 단순히 압축하려는 것으로 필터링된 후의 스트림 데이터의 저장이나 분석을 통한 활용은 고려하지 않고 있다.

비즈니스 데이터의 경우는 데이터를 다차원적으로 분석하여 온라인으로 신속하게 분석결과를 제공하는 고성능 OLAP (On-Line Analytical Processing) 시스템[2, 3, 4]들이 이미 많이 개발되어 사용되고 있

* 본 연구는 2004년 한국과학재단 우수여성과학자 도약지원연구사업(R04-2004-000-10149-0(2004)) 지원에 의해 수행되었음.

다. OLAP의 효율을 높이기 위해 다양한 데이터 저장구조, 인덱싱 기법, 집계 연산 기법, 질의 처리 기법들이 사용되고 있다. 그러나 이러한 기법들은 데이터의 주기적인 갱신을 가정하고 있기 때문에 본 연구에서 다루는 건강 관련 데이터와 같은 스트림 데이터를 저장하고 분석하기에는 적절하지 않다.

본 논문에서는 건강 데이터 수집 센서로부터 발생하는 스트림 데이터의 다차원적 분석을 위해 데이터 특징을 분석한 후 필요한 저장구조의 조건들을 제시하였고 그러한 조건들에 적절한 새로운 저장구조 모델을 제안하였다. 본 논문의 구성은 다음과 같다. 2절에서는 건강 관련 스트림 데이터에서 예외 데이터를 분석하고 필요한 저장구조의 요건을 제시한다. 3절에서는 기존 의 저장구조의 문제점을 분석한다. 4절에서는 본 논문에서 제안하는 스트림 데이터의 저장구조를 설명하고, 5절에서 제안한 저장구조의 효율성을 검증한 후 6절에서 결론을 맺는다.

2. 건강 데이터 분석

건강 관리 시스템에서 환자들을 모니터링하기 위해 추출하는 예외 데이터는 정상 데이터와 분리되는 특별한 데이터이다. 일정한 시간 간격 없이 임의로 발생하고, 발생 빈도는 대체로 매우 희박하다고 볼 수 있으며, 스트림의 형태를 갖는다. 맥박, 혈압, 체온, 혈당 등을 수집하는 센서들로부터 입력되는 예외 스트림 데이터를 질병의 종류, 투약 처방, 식사, 운동 요법 등의 정보들과 결합하여 분석을 위한 데이터를 얻어낼 수 있는데, 이러한 데이터는 다음과 같은 특징을 갖는다.

- 예외 데이터는 고차원 데이터가 될 수 있다.
- 차원들 중에는 시간 차원과 같이 끊임없이 실시간으로 증가하는 스트림 형태의 차원과 멤버가 거의 고정적인 차원들이 존재한다.
- 특정 차원들은 차원들끼리 연관성을 가지며, 유한한 개수의 조합을 발생시킨다.
- 특정 차원들의 조합에 대한 스트림 데이터의 발생에는 연관성이 존재하며, 그 연관성에 따라 데이터의 분류가 가능하다.

이러한 스트림 데이터로부터 사용자는 '당뇨병을 가진 환자A의 지난 6개월간 월별 예외 상황의 발생 빈도를 통한 병의 진전도'라든가, '지난 한달 간 심장병 환자 x와 y들의 약의 투여에 따라 나타낸 반응이나 예외 발생 빈도를 통한 투약 처방의 효율성' 등의 질의를 요구할 수 있다. 질의 처리의 패턴을 정리하면 다음과 같다.

- 예외 데이터에 대한 포인트 질의
- 특정 차원에 대한 차원 값을 변경하였을 때의 다른 차원들에 대한 범위 질의
- 여러 차원 값들이 합쳤을 때 데이터를 결합하는 집합 연산

이와 같이 건강관리 예외 데이터의 고유한 특성과 질의 패턴을 바탕으로 한 다차원적 분석을 위해서 저장구조는 다음과 같은 조건을 갖추어야 한다.

1) 예외 데이터는 고차원이고 희박하다. 따라서 '고차원 희박 데이터에 대한 처리'가 필요하다.
2) 모니터링 시스템은 예외 데이터의 각 상황에 따른 비교 분석이나 변화 추이에 대한 분석이 중요하므로, '다차원의 범위 질의 처리'가 가능해야 한다. 또한 특정 차원의 차원 값들의 결합에 따른 데이터 자체의 결합을 표현할 수 있어야 한다.
3) 예외 데이터는 차원들이 서로 특별한 연관성을 가지며, 연관성은 패턴을 형성한다. 따라서 '패턴 분석을 위한 차원 분류'가 가능해야 한다.
4) 스트림 데이터 분석에는 실시간 상태 분석이 중요하므로 '실시간 업데이트'가 가능해야 한다.
5) 저장 공간상의 제약을 극복하기 위해 '과거 데이터를 효율적으로 처리'하는 저장 구조가 필요하다.

3. 기존 저장 구조 분석

건강 스트림 데이터를 실시간에 저장하고 분석할 수 있도록 하기 위해 기존의 다차원 데이터 저장 구조와 스트림 데이터 저장 구조를 검토해 보았다. 다차원 데이터 저장 구조로는 청크 기반의 MOLAP 저장구조[2], Essbase 사의 큐브 저장구조[3], Dwarf 큐브 저장 구조[4]를 분석하였다. 스트림 데이터 저장구조로는 DSMS 시스템[1]을 살펴 보았다. 기존의 저장 구조들을 본 연구의 2절에서 제시한 기준으로 분석한 결과가 표 1에 나타나 있다.

[표1] 기존 저장구조 분석 결과

	DSMS	청크기반 MOLAP큐브	Essbase 큐브	Dwarf 큐브
고차원희박데이터	×	×	×	o
다차원범위질의	×	o	o	×
패턴분석	×	×	o	o
실시간업데이트	×	×	×	×
과거데이터처리	×	o	o	×

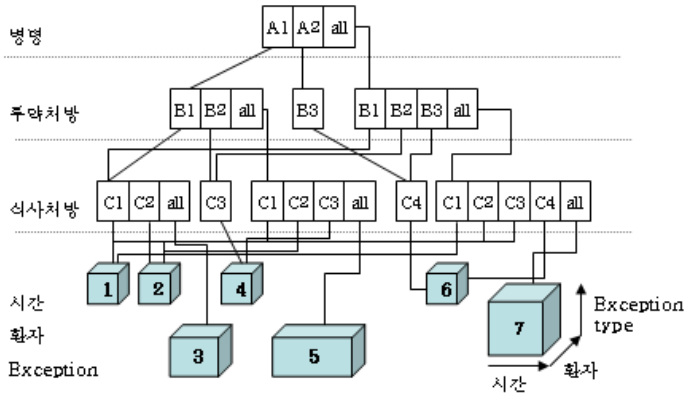
4. SMES 저장구조

기존의 OLAP 시스템들은 실시간 갱신을 필요로 하는 스트림 데이터의 처리에 한계가 있고, DSMS 시스템들은 스트림 데이터를 필터링을 통해 원하는 데이터를 추출하지만 데이터를 다차원적으로 분석하는 데 한계가 있다. 본 연구에서는 건강 관련 예외 데이터를 다차원적으로 분석하기 위해 2절에서 제시한 요구사항을 만족시키는 저장구조인 SMES (Storage for Multidimensional Analysis of Exception Stream) 저장 모델을 제안한다.

SMES는 스트림 예외 데이터 저장을 위한 요구사항을 만족하기 위해 기존의 저장 구조들을 변형하고 결합한 것이다. 차원들의 연관성을 저장 구조에 반영하고 패턴 분석이 가능하도록 데이터를 저장하기 위해 Essbase의 차원 분할 개념을 적용하여, 저장 구조를 상위 레벨과 하위 레벨로 나누었다. 상위 레벨 데이터가 고차원 희박 데이터인데 이를 효율적으로 저장하기 위해 Dwarf의 큐브 저장 구조를 적용하였다. 범위 질의, 집합 연산, 실시간 업데이트를 많이 필요로하는 하위 레벨을 저장하기 위해서는 이에 효율적인 청크 기반의 MOLAP 큐브 저장 구조를 사용하였다.

4.1 SMES 저장 모델

SMES 저장 모델은 그림 1과 같다. 그림 1은 6차원 데이터의 저장 구조이다. 상위 레벨은 병명, 투약처방, 식사처방과 같이 갱신이 자주 발생하지 않는 차원으로 Dwarf 구조로 저장된다. 상위 구조의 각 경로는 하위의 해당 MOLAP큐브로 매핑되며, MOLAP 큐브는 하위 레벨인 예외 발생 여부를 나타내는 실제 측정값이 저장된다.



[그림1] SMES 구조

1) SMES의 높이는 상위 레벨의 차원 수 + 1 이다.
2) 상위 레벨의 노드를 형성하는 셀들은 각각 [key, pointer] 값을 가지며, 각 노드는 차원 값들의 집합 연산을 위한 all 셀을 가진다.
3) 상위 레벨의 리프 노드의 포인터는 각각 그에 대응하는 MOLAP 큐브를 가리킨다.
4) 상위 레벨의 차원들은 실제 존재하는 조합만을 생성하며, 모든 차원 조합에 대한 집계 연산결과는 해당 MOLAP 큐브에서 찾을 수 있다.

4.2 차원 분류

SMES는 차원들을 특성에 따라 두 그룹으로 나눈다. 병명, 투약 처방, 식사 처방과 같이 서로 연관이 있으면서 거의 변경이 되지 않는 비스트림 데이터는 SMES의 상위 레벨에 위치시킨다. 시간, 환자, 예외 데이터 차원은 실제 입력되는 예외 데이터를 시간 단위로 저장하기 위한 것으로 이들은 SMES의 하위 레벨에 저장된다. 희박 차원도 하위 레벨에 위치시킨다. 하위 레벨의 데이터는 시간 단위로 분리하기 편하며 실시간 갱신이 효과적으로 처리될 수 있는 청크 단위의 MOLAP 큐브 구조에 저장한다. 이러한 차원 분류를 통하여 같은 종류의 예외 데이터, 즉, 같은 패턴을 가지거나 같은 경로를 가지는 데이터들이 하위의 같은 MOLAP 큐브에 모이게 되고, 질의 처리나 패턴 분석에 효율적이다.

4.3 측정값과 all 셀

SMES 하위 레벨인 MOLAP 큐브에는 실제 측정값으로 데이터 발생 여부를 나타내기 위해 비트(bit)로 데이터가 저장되고, 상위 레벨의 노드들이 차원 값 이외에 all 셀을 갖는다. 이는 가공하기 전의 실제 데이터 자체를 결합하는 집합 연산의 결과를 의미하여, Dwarf의 all 셀과는 구별되며, 다차원 슬라이스, 다이스 연산 처리를 쉽게 할 수 있도록 한다.

4.4 생성 및 업데이트 알고리즘

SMES 구조에서는 생성과 업데이트에 모두 동일 알고리즘을 사용한다. 데이터를 갱신할 때마다 SMES의 루트 노드부터 포인터를 따라 내려가다가 하위 레벨에 다다르면 해당 청크 셀의 비트를 세팅한다. 생성되지 않은 차원 값에 대해서는 새로운 노드, 셀을 생성하여 키와 포인터 값을 부여한다.

4.5 질의 처리

예외 데이터의 발생여부를 측정값으로 하는 건강관리 센서 데이터의 경우, 범위 질의를 위한 슬라이스, 다이스 연산과, 차원 값들에 대한 데이터의 결합을 의미하는 집합 연산이 필요하다. SMES에서는

상위 레벨의 각 조합에 대하여 발생한 데이터를 하나의 MOLAP 큐브에 저장하고, 차원 값들의 결합에 따른 all 셀을 가지고 있으므로 이러한 연산이 효율적으로 처리될 수 있다. P는 상위레벨 개수, C_i는 i 번째 레벨에 해당하는 노드의 cardinality, $x_1, x_2, ..., x_n$ 는 범위질의가 필요한 차원들의 레벨, u는 MOLAP 큐브 접근을 위한 단위일때, 슬라이스, 다이스 연산을 위하여 일반적으로 n 차원에 접근해야 하는 노드 개수는 표 2와 같다.

[표 2] 슬라이스,다이스 연산시 접근 노드 수

슬라이스, 다이스 연산 수행할 차원의 위치	접근해야 하는 노드 개수
상위레벨	$C_{x_1} \times C_{x_2} \times ... \times C_{x_n} \times u$
MOLAP 큐브	$P + u$
상위레벨+MOLAP큐브	$C_{x_1} \times C_{x_2} \times ... \times C_{x_n} \times u$

5. SMES 특징 및 효율성 분석

이제 SMES가 2절에서 제시한 스트림 데이터의 다차원적 분석을 위한 저장구조 요건에 적합한가를 분석해 보기로 한다.

1) 고차원 희박 데이터를 위한 저장 공간

고차원 희박 데이터를 일반적인 OLAP 저장구조에 저장하기는 힘들다. SMES에서는 차원 개수를 줄이기 위해 저장 구조를 상위 레벨과 하위 레벨로 분산하여 데이터를 저장한다. 차원 값에 변화가 적은 부분을 Dwarf 구조를 사용하여 저장, 관리한다. 차원 분리를 통해 남은 차원의 데이터 밀집도를 상대적으로 높였고, 특히 시간 차원에 대한 데이터의 밀집도가 높아지므로 희박 청크들에 대한 오버헤드의 문제는 해결되고, 이미 입증된 MOLAP 큐브의 저장 공간 효율성을 이용할 수 있다.

2) 다차원 범위질의처리

SMES는 차원 값의 결합에 따른 데이터의 집합연산의 결과인 all 셀을 사용한다. 이를 통해 슬라이스, 다이스 연산을 처리할 때, 모든 차원 값에 대하여 뿔뿔이 흩어진 데이터에 접근할 필요 없이 하나의 MOLAP 큐브만으로 연산이 가능하다.

3) 패턴분석을 위한 차원분류

SMES구조는 차원 분할 개념을 사용하는 Essbase 저장 구조와 동일한 데이터를 한번만 저장하는 Dwarf 저장 구조를 이용하여 건강 데이터를 저장한다. 이를 통해 동일한 패턴을 갖는 데이터들이 하나의 MOLAP 큐브에 집적되어 질의 처리를 효과적으로 수행할 수 있다.

4) 실시간 업데이트

스트림 데이터가 입력될 때마다 SMEMS는 큐브의 루트 노드로부터 차례대로 각 차원 값에 해당하는 노드와 셀을 따라 내려오다가 포인터가 가리키고 있는 MOLAP 큐브의 해당 셀에 데이터의 발생을 알리는 비트를 세팅한다. 해당 셀이 존재하지 않으면 노드나 셀을 생성함으로써 실시간 업데이트를 가능하게 하였다.

5) 과거데이터의 처리

시간이 많이 경과된 스트림 데이터는 상위 레벨의 집적된 데이터만을 제외하고 폐기 처분하거나 고도로 압축해야 한다. SMES는 하위의 MOLAP 큐브에 sliding window를 적용하여 최근 데이터만을 유지하고, 시간 차원에 대해서는 시간 계층 구조를 써서 데이터를 압축하는 Tilt Time 기법을 적용하여 오래된 데이터일수록 간결하게 압축하도록 하였다.

6. 결론

본 논문에서는 센서로부터 입력되는 건강 스트림 데이터를 다차원적으로 분석하기 위한 저장 구조를 제안하였다. 이러한 데이터의 저장 구조가 갖춰야 할 조건들을 분석하였고, 이를 만족하는 저장구조인 SMES를 제안하였다. 최근 유비쿼터스 컴퓨팅 환경이 구축되면서 스트림 형태의 데이터를 신속하게 분석하고 효율적으로 관리할 필요성이 대두되고 있다. 본 연구는 SMES 구조를 통해 건강 스트림 데이터 뿐 아니라 각종 모니터링 시스템으로부터 발생하는 스트림 데이터의 저장, 분석에 유용한 저장 구조 모델을 제공하였다.

참고문헌

[1] Stanford Stream Data Management (STREAM) Project. http://www-db.stanford.edu/stream

[2] Yihong Zhao, Prasad Deshpande, Jeffrey Naughton, "An Array-Based Algorithm for Simultaneous Multidimensional Aggregates," Proc. ACM SIGMOD, pp. 159-170, 1997.

[3] Hyperion Corp. "Large-Scale Data Warehousing Using Hyperion Essbase OLAP Technology," http://www.hyperion.com/downloads/teraplex.pdf

[4] Yannis Sismanis, Antonios Deligiannakis, Nick Roussopoulos, Yannis Kotidis, "Dwarf: Shrinking the PetaCube," Proc. ACM SIGMOD '2002,